Person-fit analysis

Meijer, R.R.; Sijtsma, K.

Published in:
Measurement problems in social and behavioral research

Publication date:
1995

Chapter 3

PERSON-FIT ANALYSIS: CLASSIFICATION OF PERSONS ON THE BASIS OF THEIR ITEM SCORE PATTERNS

R.R. Meijer and K. Sijtsma

Several methods have been proposed for the detection of examinees that produce unusual or rare patterns of item scores in ability and achievement tests. These examinees may be of special interest as far as their behavior underlying test performance deviates from the behavior that is exhibited by most examinees and that leads to normal or common item score patterns. Thus, a low-ability examinee that produces many correct item scores on difficult items by copying them from his more able neighbor in a group-administered test session is expected to produce an item score pattern that deviates from what he would have produced had he relied solely on his own ability. In order to have a more realistic evaluation of this person's performance it would be useful to have a method that enables the researcher to detect such persons on the basis of their item score patterns.

An item score pattern may be aberrant in two different ways. First, it may deviate from the patterns produced by the majority of the population. If the item scores in this population have not been fitted to a particular model a group-based assessment of aberrance may be pursued. Second, it may be unexpected given a particular measurement model. This leads to a model-based assessment of aberrance. In this chapter, we will concentrate on model-based assessment of aberrance because it yields a firmer basis for conclusions about aberrance.

Most model-based methods for the detection of aberrant persons are based on item response theory (IRT) models. We will first discuss this class of models. Next, we will discuss the class of methods for the detection of aberrant persons known as appropriateness measurement or person-fit analysis. Finally, the results from a study about the influence of item, test, person, and group characteristics on decisions about aberrance on the basis of a nonparametric method are presented and discussed.

Item Response Theory

Item response theory offers an approach to the analysis of item scores that has several advantages over the well-known classical test theory (CTT) approach.

IRT models are defined in such a way that observable consequences can be derived from them. In other words, predictions about the structure of empirical data can be derived from IRT models. These predictions can be compared with the actual data structure. Statistical tests or descriptive methods can be used to decide whether the discrepancy between the expected data structure and the empirical data structure is small enough to accept the model.

If the model is accepted as a satisfactory explanation of the data structure the characteristics of the particular IRT model are valid for the test and the population under consideration. These characteristics are discussed next.

1. The measurement level of the scale on which persons and items are displayed can be derived from the formulation of a particular IRT model. For example, the Mokken (1971) approach to IRT yields ordinal measurement whereas the Rasch (1960) model yields a metric scale allowing interval, difference or ratio measurement.

CTT assumes that persons can be measured on an interval scale. However, the classical model describes the error component in measures of abilities and traits but does not restrict the data by imposing a particular model structure. For example, the requirement of a high reliability for test scores does not logically follow from CTT. This is rather a desideratum that is formulated by measurement practitioners. Thus fit of a 'classical model' to the data is not an issue in the context of CTT. The correctness of the assumption of interval measurement thus can not be checked. As a result, the interval scale from CTT does not have an empirical foundation (Torgerson 1958).

2. Several IRT models allow the item-free measurement of persons, the person-free measurement of items, or both. Item-free measurement of persons is explained by means of an example. Suppose that the performance of Dutch pupils from primary education on the items from a pool measuring spatial orientation can be explained by the Rasch (1960) model. Item-free measurement means that a person's ability parameter can be estimated by means of any subset of spatial orientation items from the item pool irrespective of the difficulty level of the items used. Thus, spatial orientation tests of low, medium, and high difficulty level can be used to estimate the same ability parameter. Only the accuracy of the estimates is affected by the particular choice of items. Person-free measu-

rement of items has a similar interpretation.

It may be noted that, for example, measurement of ability in a CTT context is highly dependent on the choice of the items. A person with a medium ability in spatial orientation has a high true score on an easy test, an average true score on a medium difficulty test, and a low true score on a difficult test. Not only is the accuracy of parameter estimation affected by the choice of the items, each test from the same item pool estimates a different ability true score parameter. As a result, the comparison of test performances of different examinees on tests of different difficulty levels becomes problematic. For example, if examinee a has a low true score on a difficult test and examinee b has an average true score on a medium difficulty test, which of the two has the better ability of spatial orientation? By means of, for example, the Rasch model the ability parameters can be estimated independent of the difficulty levels of both tests and a direct comparison of these two examinees is thus feasible.

3. For some IRT models the accuracy of the estimate of the ability parameter can be estimated as a function of the scale and of the items used for measurement. Thus, given a set of items with known characteristics two examinees with different abilities are measured with different degrees of accuracy. In particular, the examinee whose ability level matches the difficulty level of the test best is measured the most accurately.

Note that CTT allows the assessment of accuracy or reliability of measurement as a group characteristic but not on the individual level. For practical applications it is thus assumed that the same reliability and the same standard measurement error hold for each member of the group.

Next, a few well-known IRT models relevant to person-fit analysis are presented. Each model is discussed in the light of the general characteristics of IRT models discussed previously.

Item Response Models

The IRT models that are relevant for this chapter assume that 1) a test is unidimensional in the sense that all items measure one common ability or trait which is represented by the latent variable Θ; 2) the probabilities of giving the correct answer to single items are locally independent, that is for given Θ the probability of answering item g correctly is not affected by the probability of answering preceding items correctly nor does it affect probabilities for consecutive items.

A key concept in IRT is the item response function (IRF). This

function gives the relationship between the latent ability parameter Θ and the probability that given Θ a correct response is given to a particular item g. For the class of IRT models that is relevant to appropriateness measurement the IRF is assumed to be monotonely nondecreasing. Thus, the higher the ability (Θ) of spatial orientation, the greater the probability that a correct answer is given to an item for the measurement of spatial ability.

IRT models differ in the restrictions that are placed on the IRFs. An important distinction can be made between parametric and nonparametric IRT models. In parametric models the IRFs are defined by parametric functions that allow the exact determination of success probabilities provided that parameters of persons and items (such as their abilities and difficulties, respectively) are known. In nonparametric models the IRFs are subjected to order restrictions such as: the higher Θ the greater the success probability. Although such order restrictions lead to empirically useful models they do not allow the exact determination of the success probability.

Parametric models thus lead to more information about examinees and items than nonparametric models. The price to be paid for this is that, in general, parametric models fit the data less easily than nonparametric models. Put differently, several parametric models impose more structure on the data than nonparametric models and this may lead to shorter tests with a less reliable test score. The choice of a particular model may depend on the application of the test envisaged; refer to Meijer, Sijtsma, and Smid (1990) for further discussion.

The Rasch Model

Let Θ denote the ability parameter and let δ_g denote the difficulty parameter of item g. Further, let X_g denote the item score variable with realizations 1 for a correct answer and 0 for an incorrect answer. Finally, P denotes a probability and exp(.) the exponential function. Then the IRF of the one-parameter logistic Rasch (1960) model is defined as

$$P(X_g = 1|\theta) = \frac{\exp(\theta - \delta_g)}{[1 + \exp(\theta - \delta_g)]}.$$

Items are allowed to have different difficulties (locations on the Θ scale) but apart from that they must have the same psychometric properties.

Provided that the Rasch model fits the data the following characteristics are obtained. Both persons and items are measured on a metric scale

which can be defined as interval, difference or ratio depending on the exact formulation of the model. For the formulation presented here a difference scale is obtained. Measurement of persons and items can be characterized as item-free and person-free, respectively. In the context of the Rasch model these characteristics are known as population independent person and item measurement, respectively. The accuracy of parameter estimation is a function of the latent θ scale and the items used for measurement.

Other Parametric Models

Less restrictive models are the two-parameter logistic and the three-parameter logistic Birnbaum (1968) models. In addition to difficulty parameters for each item, these models include a discrimination parameter for each item, and a discrimination parameter and a pseudo-chance or guessing parameter for each item, respectively. A detailed introduction to these models would be beyond the scope of this chapter.

The Mokken Models

Two well-known nonparametric IRT models have been proposed by Mokken (1971) and Mokken and Lewis (1982). The model of monotone homogeneity stipulates the IRFs to be monotonically nondecreasing functions of the latent person variable θ. In particular, for two persons i and j,

$$\textit{if } \theta_i < \theta_j \textit{ then } P(X_g = 1|\theta_i) \leq P(X_g = 1|\theta_j), \textit{ all } i,j.$$

Note that each item orders persons identically. Therefore, each subset of items also orders persons identically. Thus, the person ordering is item-free. The model of monotone homogeneity leads to item-free person measurement on an ordinal scale. A person-free ordering of items can not be realized because the IRFs are allowed to intersect and, as a result, the ordering of success probabilities depends on θ. Measurement precision is estimated for the number-correct score in a particular population. In practice, it has to be assumed that all persons are measured equally accurately.

In addition, Mokken (1971) and Mokken and Lewis (1982) proposed the model of double monotonicity. This is a more restrictive model than the first model because it adds the assumption that the IRFs do not intersect. This extra assumption says that the ordering of the items according to their difficulty is the same for each person. Let us assume

that k dichotomously scored items have been numbered and ordered such that

$$P_1(\theta) \geq P_2(\theta) \geq \ldots \geq P_k(\theta), \textit{ with } P_g(\theta) = P(X_g = 1|\theta), \textit{ all } \theta.$$

Note that except for ties this ordering holds for all persons. Thus, item ordering is person-free. Because the model of double monotonicity is a special case of the model of monotone homogeneity an item-free ordering of persons is also realized.

Aberrant Item Score Patterns

Let us assume the item ordering proposed at the end of the preceding section. The first item thus is the easiest for each person, and so forth. Before we go on with the presentation of a few examples of aberrant item score patterns it may be noted that the assumption of an identical item ordering for each person is not necessary in person-fit analysis. For example, several methods have been proposed in the context of the two- and three-parameter logistic models that allow IRFs to cross. However, the assumption of an identical item ordering greatly facilitates the interpretation of results.

Several examples of aberrant behavior leading to aberrant item score patterns can be given. We will give three. We will assume unidimensionality, local independence, and nondecreasing and nonintersecting IRFs. Furthermore, we assume that item responses are the result of a stochastic process which excludes success probabilities of 0 or 1. With this assumption each item score pattern has a probability greater than 0. However, several patterns are very unlikely given the set of assumptions and may have been produced as a result of other mechanisms than the ability that the test intends to measure.

Given the set of assumptions which includes a fixed item ordering it is expected that persons will often obtain 1s on the first few items and 0s on the last few items. The higher the ability Θ the more 1s and the less 0s are expected. For some persons the test may provoke behavior that results from abilities or traits $\Theta^{(1)}$, $\Theta^{(2)}$, ..., in addition to Θ. As a result, the expected pattern of 0s and 1s may not be produced but rather patterns that are very unlikely given our set of assumptions which includes unidimensional measurement.

For example, in a group-administered test an examinee with a low ability Θ may decide to sit next to a high-ability student and copy the answers to the most difficult items from him or her. This might result in a pattern of item scores with relatively many 1s for the easiest and the

most difficult items, and relatively few 1s for the items of medium difficulty. An example of such a pattern would be (k = 15)

(110110010000101).

Another example is that an examinee with a low ability guesses for the correct answer on many of the multiple-choice items with a constant success probability of A^{-1} (A alternatives). This strategy is expected to result in an item score pattern that correlates 0 with the item difficulty, for example

(010100010010100).

The final example entails an examinee with a very high ability who produces correct answers on almost all items except the two easiest. For him or her these easiest items are almost too simple to be true. Therefore, these items are reinterpreted and, consequently, a correct answer is given to a question that was not asked. This may result in an item score pattern like

(001111111111111).

Note that aberrant behavior need not result in an aberrant item score pattern, for example, if someone who cheats copies his or her answers from a neighbor that has a low ability. In addition, an unlikely item score pattern does not always unequivocally reveal the cause of aberrance. For example, the pattern that was presented as a typical result of guessing could, in absence of further evidence, also be interpreted as the result of cheating by a low-ability examinee. Thus if aberrance is observed, additional information is needed to produce a correct interpretation.

Person-Fit Statistics

Many statistics have been proposed that express the degree of aberrance of item score patterns. These statistics can be divided into three classes (Meijer, 1994, chap. 1).

Parametric Person-Fit Statistics

The first class consists of statistics proposed in the context of parametric IRT models. Several statistics are based on the likelihood of the item score pattern under a particular IRT model given the maximum likelihood estimate of $\hat{\theta}$ denoted by $\hat{\theta}$; for an arbitrarily chosen person i:

$$L_i = \sum_g \left\{ X_{ig} lnP(\hat{\theta}_i) + (1 - X_{ig}) \ln[1 - P_g(\hat{\theta}_i)] \right\}.$$

The L statistic is due to Levine and Rubin (1979). Because L is expressed on a logarithmic scale, obviously $L < 0$. L has large negative values if $X_{ig} = 0$ while $P_g(\hat{\theta}_i)$ is large and if $X_{ig} = 1$ while $P_g(\hat{\theta}_i)$ is small, for a relatively large numbers of items. Other proposals, not based on the log likelihood, have been done by, for example, Tatsuoka (1984).

Nonparametric Person-Fit Statistics

The second class of statistics (for example, Van der Flier, 1980, 1982; Rosenbaum, 1987) takes as given a nonparametric IRT model. Van der Flier (1980, 1982) proposed the U3 statistic that was based on the nonintersection of IRFs. Let $\mathbf{X}$, $\mathbf{X}^*$, and $\mathbf{X}'$ denote three vectors with k dichotomous item scores (for example, let k = 15, and X = 8):

$$\boldsymbol{X} = (X_1, \ldots, X_k); \textit{ for example, } (1101111100110000);$$

$$\boldsymbol{X}^* = (X_1^*, \ldots, X_k^*); \textit{ for example, } (1111111110000000);$$

$$\boldsymbol{X}' = (X_1', \ldots, X_k'); \textit{ for example, } (0000000111111111).$$

Let the first vector be the observed item score vector for given number-correct score X, then the second vector is the item score pattern expected under the deterministic Guttman (1950) model, and the third vector is the vector least probable given the Guttman model. For an arbitrarily selected person i the U3 statistic is defined as

$$U3_i = \frac{lnP(X_i^*) - lnP(X_i)}{lnP(X_i^*) - lnP(X_i')}.$$

Note that $0 \leq U3 \leq 1$. Further, if $\mathbf{X} = \mathbf{X}^*$ then U3 = 0, and if $\mathbf{X} = \mathbf{X}'$ then U3 = 1. Van der Flier (1980) showed that for given X, U3 is an increasing function of the number of item pairs in which the easiest item was answered incorrectly and the more difficult item was answered correctly. A theoretical sampling distribution has been derived for U3. This greatly facilitates the interpretation of numerical values of this statistic.

Group-Based Person-Fit Statistics

The third class of statistics is not based on any IRT or other model but on group characteristics (refer to Meijer 1994, chap. 1, for an overview). A statistic from this class is Sato's (1975) Caution Index C. Let **n** be the vector containing the frequencies n_g (g = 1, ..., k) of correct answers on the items. Furthermore, let $\sigma(.)$ be the within-person covariance between the elements of two vectors. For an arbitrarily chosen person i the Caution Index is defined as

$$C_i = 1 - \frac{\sigma(X_i,n)}{\sigma(X_i^*,n)}.$$

It may be noted that $C \geq 0$. The minimum of 0 is obtained if $\mathbf{X} = \mathbf{X}^*$. If $\sigma(\mathbf{X}_i,\mathbf{n}) < 0$ then $C > 1$. Thus relatively large values of C indicate aberrance. However, because the sampling distribution of C is unknown the interpretation of the numerical values may be problematic.

Factors Affecting the Power of a Person-Fit Statistic

Little is known about the factors that might influence the power of person-fit statistics to detect aberrant persons by means of their item score pattern (for example, Reise and Due 1991). Meijer, Molenaar, and Sijtsma (1993) studied the influence of characteristics of the items, the test, the individuals, and the population on the power of the U3 statistic (Van der Flier 1980, 1982). These characteristics were 1) the discrimination power of the items denoted α; 2) the test length; 3) the aberrance or normality of an individual; and 4) the ratio of aberrants to normals in the group.

Method

The design for the power study was a completely crossed four-factorial 4 x 2 x 2 x 2 design. The Item Factor had 4 levels of uniform discrimination power for all items in the test, ranging from weak to strong discrimination: α = .5, 1, 2, and 5. The Test Factor had 2 levels corresponding with relatively short and long tests: k = 17 and 33. The Individual Factor had 2 levels of aberrant behavior: an aberrant person either cheated on the most difficult items to obtain correct answers or guessed for the correct answer with a constant success probability for all items despite their difficulty. The Group Factor had 2 levels: a sample (size: 450) contained 5.5 (25 out of 450) or 11 percent (50 out of 450) of aberrant persons.

Data matrices of the order n x k (n = 450; k = 17 or 33) containing binary scores were simulated by means of the two-parameter logistic model (Birnbaum, 1968) and a standard normal distribution of θ. Guessing was simulated by assuming a constant success probability of .25 on each item irrespective of the difficulties. Cheating was simulated by assuming that 1) the ability was below average ($\theta < 0$), that 2) cheating only occurred on the 3 (k = 17) or 6 (k = 33) most difficult items, and that 3) cheating always resulted in a correct answer (item score 1).

Within each cell of the design 8 independent samples were drawn to assess the stability of the results across independent repetitions.

For each sample the aberrant simulees were known. The sample U3 values were ordered according to increasing magnitude. The 25 or 50 simulees with the highest sample U3 values were identified and the percentages of valid aberrants (correctly identified aberrants), valid normals (correctly identified normals), false aberrants (normals mistakenly identified as aberrants), and false normals (aberrants mistakenly identified as normals) was determined. For each cell the mean percentages across 8 independent repetitions were reported.

In addition, it was investigated to what degree results could be repeated across two independent repetitions. Specifically, the percentage of aberrants successfully detected by means of U3 in two samples was determined and the mean percentage across 4 pairs of samples was reported. Similar results were obtained for valid aberrants, valid normals, false aberrants, and false normals.

Results

A limited part of the results from this research is tabulated here. Refer to Meijer et al. (1993) for more detailed results.

From Table 1 it can be concluded that an increase in the discrimination power, the test length, and the percentage of aberrants each led to a higher percentage of correct classification of cheating aberrants. The results for the Guessing Condition are highly similar. Results for valid normals, false aberrants, and false normals are not provided because they can be deduced from Table 1 in the following way.

Since the marginals of the two-by-two tables of simulated normals and aberrants (marginals are 400 and 50, and 425 and 25, respectively) by empirically classified simulees (400 low and 50 high U3 values, and 425 low and 25 high U3 values, respectively) are equal, the numbers of false normals and false aberrants are equal. Given these fixed and equal marginal distributions it can easily be verified that if the percentage of valid aberrants increases, then the percentage of valid normals also

increases and the percentages of false aberrants and false normals decrease.

Table 1. **Mean and Standard Deviation (SD) of the Percentage of Cheating Simulees Classified as Valid Aberrant Averaged across Eight Replications.**

NA = 50

k		17		33	
	α				
		Mean %	*SD*	*Mean %*	*SD*
	.5	55.8	4.1	74.8	3.6
	1.0	73	5.9	88.3	2.5
	2.0	90	3.4	97	1.7
	5.0	99.3	1.0	100	0

NA = 25

k		17		33	
	α				
		Mean %	*SD*	*Mean %*	*SD*
	.5	40.5	8.1	64.5	5.5
	1.0	57	8.8	81.5	8.0
	2.0	88	4.9	94.5	2.8
	5.0	100	0	100	0

Note: NA = number of aberrants; k = number of items; α = discrimination power.

From Table 2 it can be concluded that the percentage of replicable cheating simulees classified as aberrant also increases with an increase in the discrimination power, the test length, and the percentage of aberrants. The results for the Guessing Condition are highly comparable. Note that compared with Table 1 the percentages are smaller in Table 2 because a percentage on the basis of two repetitions can not exceed the smallest of the two percentages on the basis of separate samples.

Refer to Meijer et al. (1993) for detailed results concerning the percentages of cheating and guessing simulees detected in one of the two replications only. Here these results are summarized by noting that, in general, these percentages decrease as the discrimination power increases, the test length increases, and the percentage of aberrants increases.

Finally, in all cells of the design a high level of agreement of U3 values between two repetitions was found.

Table 2. **Mean and Standard Deviation (SD) of the Percentage of Replicable Cheating Simulees Classified as Aberrant Averaged across Four Pairs of Replications.**

NA = 50

k	17		33	
α				
	Mean %	*SD*	*Mean %*	*SD*
.5	25	8.1	54.5	3.6
1.0	53.5	6.7	79	2.4
2.0	80.5	2.2	94	2.8
5.0	98.5	0.9	100	0

NA = 25

k	17		33	
α				
	Mean %	*SD*	*Mean %*	*SD*
.5	18	4.5	42	6.0
1.0	37	8.7	68	8.5
2.0	80	2.8	89	3.3
5.0	100	0	100	0

Note: NA = number of aberrants; k = number of items; α = discrimination power.

Discussion

In general, person-fit analysis will be used as an exploratory technique to find persons who behave unexpectedly on the basis of an IRT model or with respect to the other persons in the group. In the relatively rare cases in which the researcher expects a particular kind of aberrant behavior, it is advisable to construct tests in which the most difficult items provoke aberrant responses. Aberrant behavior may be difficult to recognize ad hoc if the items used are not specifically tuned to elicit exactly a particular kind of aberrant behavior. Refer to Frary (1993) for a discussion of a class of statistics that were designed especially to detect cheating.

Reise and Due (1991) concluded that tests used for person-fit analysis should be long and the item difficulties should have a large dispersion, that is, the test should contain items of low, medium, and high difficulty. Meijer et al. (1993) concluded that person-fit analysis is more effective with increasing test length, increasing discrimination power, and an increasing percentage of aberrants in the population. Holding everything else constant, an increasing discrimination power is identical to an increasing dispersion of the item difficulties (Mokken, Lewis, and Sijtsma 1986). The effects of an increasing percentage of aberrants in the population is also known in the context of personnel selection where a larger base rate implies a larger success ratio (Wiggins 1973).

Fortunately, there is a trade-off between test length and discrimination power. Thus short tests may be used for person-fit analysis provided that the items have strong discrimination power. Likewise if items have weak discrimination power this can be compensated for by a large number of such items. For example, if the discrimination power is 1.0 and the test length is 33, the percentage of valid aberrants is about the same as for a discrimination power of 2.0 and a test length of 17 (Table 1; Cheating Condition, base rate is .11, θ standard normally distributed). Given a standard normal distribution of θ a discrimination parameter of 1.0 can be considered moderate whereas a discrimination parameter of 2.0 is relatively high (Hambleton and Swaminathan 1985, p. 36), but not unrealistic.

REFERENCES

Birnbaum, A. (1968),
'Some latent trait models and their use in inferring an examinee's ability', in: Lord, F.M. and F.M. Novick, *Statistical theories of mental test scores,* Reading MA (Addison-Wesley).

Flier, H. van der (1980),
Vergelijkbaarheid van individuele testprestaties (*Comparability of individual test performance*), Lisse (Swets and Zeitlinger).

Flier, H. van der (1982),
'Deviant response patterns and comparability of test scores', *Journal of Cross-Cultural Psychology*, 13, p. 267-298.

Frary, R.B. (1993),
'Statistical detection of multiple-choice answer copying: review and commentary', *Applied Measurement in Education*, 6, p. 153-165.

Guttman, L. (1950),
'The basis for scalogram analysis', in: Stouffer, S.A., L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star and J.A. Clausen (eds), *Measurement and prediction*, Princeton (Princeton University Press), p. 60-90.

Hambleton, R.K. and H. Swaminathan (1985),
Item response theory. Principles and applications, Boston (Kluwer-Nijhoff).

Levine, M.V. and D.B. Rubin (1979),
'Measuring the appropriateness of multiple-choice test scores', *Journal of Educational Statistics*, 4, p. 269-290.

Meijer, R.R. (1994),
Nonparametric person fit analysis, Vrije Universiteit Amsterdam (unpublished dissertation).

Meijer, R.R., I.W. Molenaar and K. Sijtsma (1994),
'Influence of test and person characteristics on nonparametric appropriateness measurement', *Applied Psychological Measurement*, 18 (in press).

Meijer, R.R., K. Sijtsma and N.D. Smid (1990),
'Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT', *Applied Psychological Measurement*, 14, p. 283-298.

Mokken, R.J. (1971),
A theory and procedure of scale analysis, The Hague (Mouton).

Mokken, R.J. and C. Lewis (1982),
'A nonparametric approach to the analysis of dichotomous item responses', *Applied Psychological Measurement*, 6, p. 417-430.

Mokken, R.J., C. Lewis and K. Sijtsma (1986),
'Rejoinder to "The Mokken scale: A critical discussion"', *Applied Psychological Measurement*, 10, p. 279-285.

Rasch, G. (1960),
Probabilistic models for some intelligence and attainment tests, Copenhagen (Nielsen and Lydiche).

Reise, P.R. and A.M. Due (1991),
'The influence of test characteristics on the detection of aberrant response patterns', *Applied Psychological Measurement*, 15, p. 217-226.

Rosenbaum, P.R. (1987),
'Probability inequalities for latent scales', *British Journal of Mathematical and Statistical Psychology*, 40, p. 157-168.

Sato, T. (1975),
The construction and interpretation of S-P tables, Tokyo (Meiji Tosho (in Japanese)).

Tatsuoka, K.K. (1984),
'Caution indices based on item response theory', *Psychometrika*, 49, p. 95-110.

Torgerson, W. (1958),
Theory and methods of scaling, New York (Wiley).

Wiggins, J.S. (1973), *Personality and prediction. Principles of personality assessment*, Reading MA (Addison-Wesley).